BUILDING SEARCH ENGINE USING MACHINE LEARNING

K. VENKATESH, VENDRA SITA MAHALAKSHMI

PG student, D.N.R. COLLEGE, P.G. COURSES (AUTONOMOUS), BHIMAVARAM-534202, AP.
E mail id:- vsmlakshmi444@gmail.com
Assistant Professor in DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS, BHIMAVARAM-534202, AP.
E mail id:- kornalavenkatesh@gmail.com

ABSTRACT

The web is the huge and most extravagant wellspring of data. To recover the information from the World Wide Web, Search Engines are commonly utilized. Search engines provide a simple interface for searching for user query and displaying results in the form of the web address of the relevant web page, but using traditional search engines has become very challenging to obtain suitable information. This paper proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.

1 INTRODUCTION

World Wide Web is actually a web of individual systems and servers which are connected with different technology and methods. Every site comprises the heaps of site pages that are being made and sent on the server. So if a user needs something, then he or she needs to type a keyword. Keyword is a set of words extracted from user search input. Search input given by a user may be syntactically incorrect. Here comes the actual need for search engines. Search engines provide you a simple interface to search user queries and display the results.

1) Web crawler Web crawlers help in collecting data about a website and the links related to them. We are only using web crawlers for collecting data and information from WWW and storing it in our database.

2) Indexer Index which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository.

3) Query Engine It is mainly used to reply to the user's keyword and show the effective outcome for their keyword. In the query engine, the Page ranking algorithm ranks the URL by using different algorithms in the query engine.

4)This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the PageRank algorithm is given as input to the machine learning algorithm.

5)The section II discusses the related work in search engine and PageRank algorithm. In section III Objective is explained. Section IV deals with a proposed system which is based on machine learning technique and section V contains the conclusion.

2. LITERATURE SURVEY AND RELATED WORK

2.1 Weighted page rank algorithm based on in-out weight of webpages

AUTHORS: Kalyani Desikan, B. Jaganathan.

In its classical formulation, the well known page rank algorithm ranks web pages only based on in-links between web pages. We propose a new in-out weight based page rank algorithm. In this paper, we have introduced a new weight matrix based on both the in-links and out-links between web pages to compute the page ranks. We have illustrated the working of our algorithm using a web graph. We notice that the page rank values of the web pages computed using the original page rank algorithm and our proposed algorithm are comparable. Moreover, our algorithm is found to be efficient with respect to the time taken to compute the page rank values.

2.2Web Page Ranking Using Machine Learning Approach

AUTHORS: Junaid Khan, Arunima Jaiswal.

One of the key components which ensures the acceptance of web search service is the web page ranker - a component which is said to have been the main contributing factor to the early successes of Google. It is well established that a machine learning method such as the Graph Neural Network (GNN) is able to learn and estimate Google's page ranking algorithm. This paper shows that the GNN can successfully learn many other web page ranking methods e.g.

Trust Rank, HITS and OPIC. Experimental results show that GNN may be suitable to learn any arbitrary web page ranking scheme, and hence, may be more flexible than any other existing web page ranking scheme. The significance of this observation lies in the fact that it is possible to learn ranking schemes for which no algorithmic solution exists or is known.

2.3Review of features and machine learning techniques for web searching.

AUTHORS: Neha Sharm ,Narendra Kohli

As the amount of information is growing rapidly on world wide web, it has become very difficult to get relevant information using traditional search engines within a stipulated time. The main reasons for irrelevant search results are the lack of understanding of user's search intention or user's preferences, keyword based searching, short queries. In this paper, we will study different features that are used in information retrieval. We will also discuss various machine learning techniques that are helpful in deciding the relevance of web page to user. We have done classification on the basis of features. In the end we will compare different techniques and their pros and cons are also discussed.

3 EXISTING SYSTEM

The current approach to building search engine using machine learning involves collecting and preprocessing data, extracting useful features, and employing machine learning models for ranking relevance. user queries are processed, and user –friendly interface displays result. Continuous improvement, user feedback integration, scalability consideration, and data privacy measures are integral to the existing system.

DISADVANTAGES:

- DATA REQUIREMENT: -Quality training data is time-consuming and costly.
- RESOURCEINTESIVE: -Machine learning can demand powerful hardware.
- CONTINUOUS TRAINING: -Regular updates are needed.
- LACK OF TRANSPARENCY: -Some models are not easily explainable.
- USER FEEDBACK CHALLENGES: -Not all users provide feedback.

4 PROPOSED WORK AND ALGORITHM

Integrate data collection, preprocessing indexing, feature extraction, and a ranking algorithm to create an efficient search engine. Enhance result with machine learning models, and continuously refine based on user feedback, scalability considerations, user interface design, and continuous model improvement.

ADVANTAGES:

- DATA COLLECTION: -Gather diverse data for indexing.
- QUERY UNDERSTANDING: -Implement a mechanism to understand user queries better.
- SECURITY: -Ensure data privacy and protection.
- PREPROCESSING: -Clean and structure data.
- USER INTERFACE: -Create user-friendly design.

5 METHODOLOGIES

MODULES

- Manager
- user
- Admin
- Machine-learning

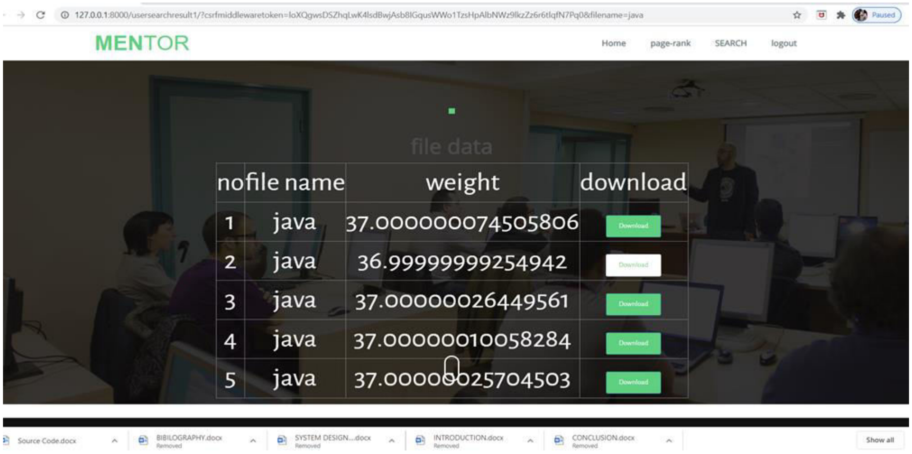
Manager: Manager information and task description for the entire experiment.

User: user information and task descriptions for the entire experiment.

user after login into the session he will get two options. he can search the whatever particular URL or information. we can search the particular file and also we canget the weight and rank of the file by using the page rank.

.**Admin:** Admin will give authority to managers and users. In order to facilitate activate the managers and activate the users. the admin can see the details of allusers and managers. Admin can get the accuracy results of SVM and XGBOOST algorithms.

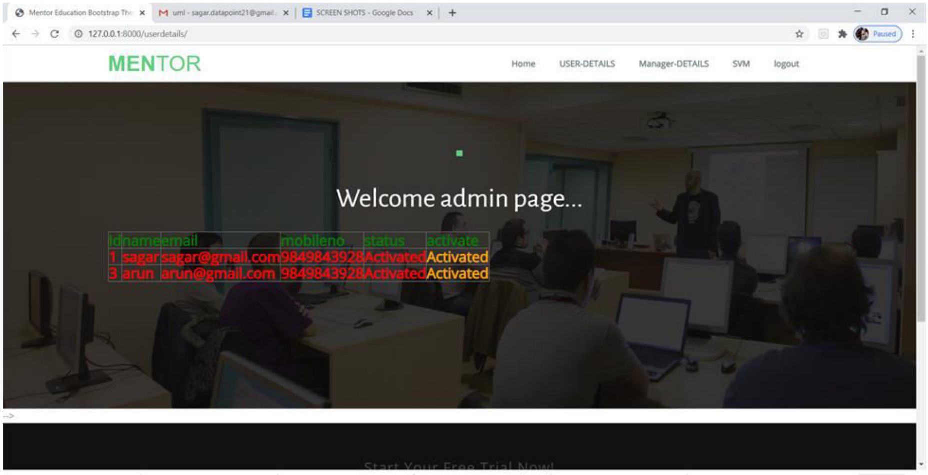
6 RESULTS AND DISCUSSION



file data

no	file name	weight	download
1	java	37.000000074505806	Download
2	java	36.99999999254942	Download
3	java	37.00000026449561	Download
4	java	37.00000010058284	Download
5	java	37.00000025704503	Download

Fig1: User Details

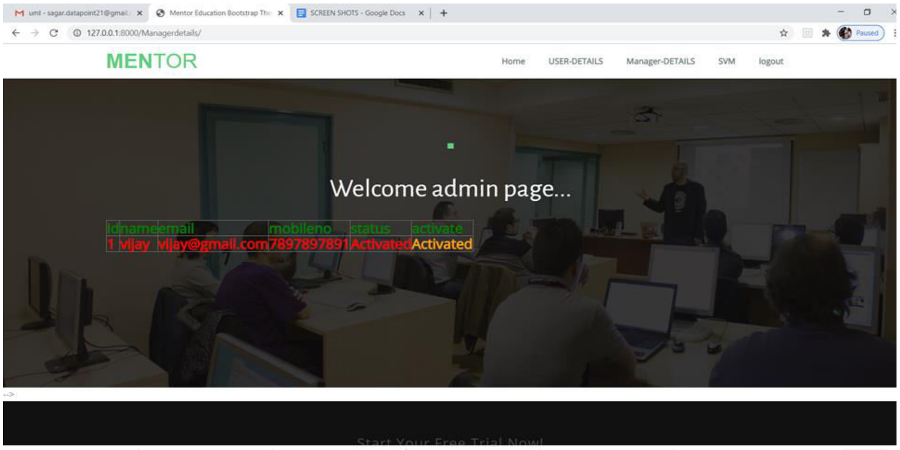


Welcome admin page...

Id	name	email	mobileno	status	activate
1	sagar	sagar@gmail.com	9849843928	Activated	Activated
3	arun	arun@gmail.com	9849843928	Activated	Activated

Start Your Free Trial Now!

Fig 2: Welcome Admin Page

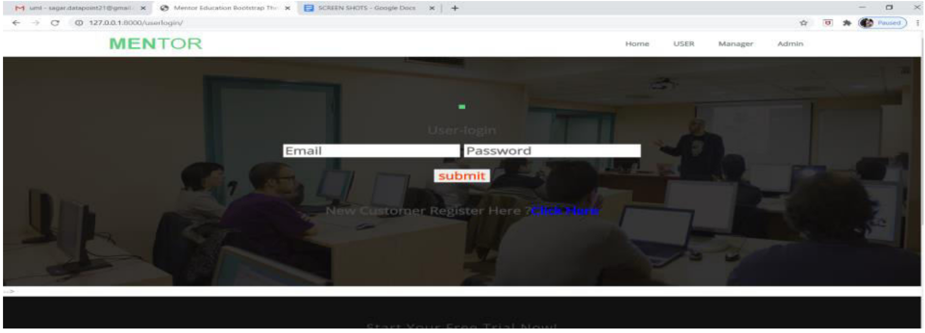


Welcome admin page...

Id	name	email	mobileno	status	activate
1	vijay	vijay@gmail.com	7897897891	Activated	Activated

Start Your Free Trial Now!

Fig 3: Manager Details


uml - sagar.datapoint21@gmail
Mentor Education Bootstrap Th
SCREEN SHOTS - Google Docs
MENTOR
Home
USER
Manager
Admin
Paused
submit
Start Your Free Trial Now!

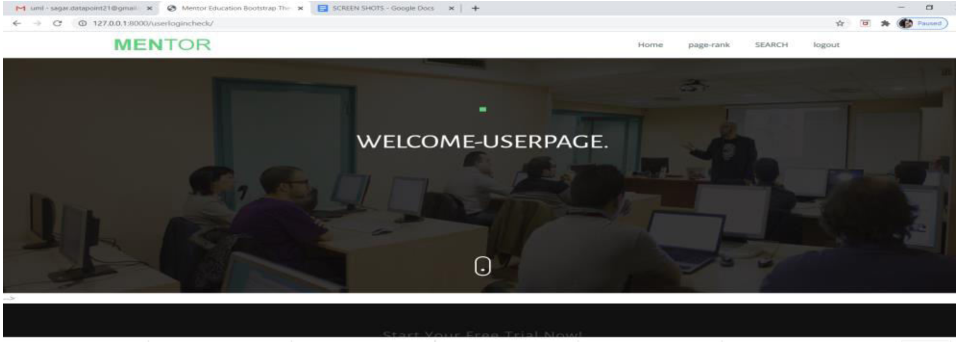
127.0.0.1:8000/userlogin/

User-login

Email Password

New Customer Register Here ?Click Here

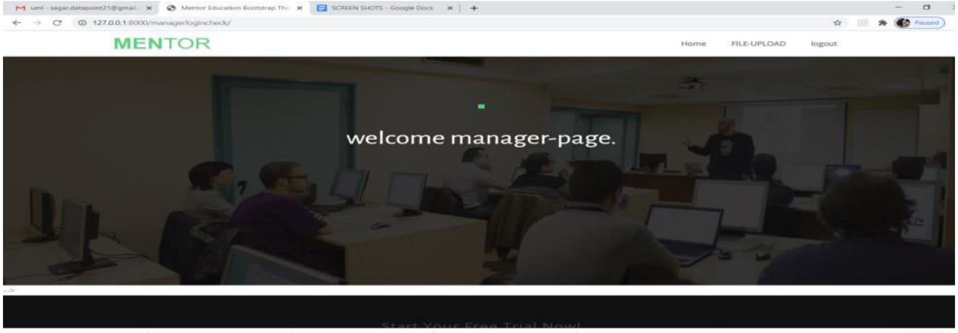
Fig 4 : User Login


uml - sagar.datapoint21@gmail
Mentor Education Bootstrap Th
SCREEN SHOTS - Google Docs
MENTOR
Home
page-rank
SEARCH
logout
Paused
Start Your Free Trial Now!

127.0.0.1:8000/userloginchecK/

WELCOME-USERPAGE.

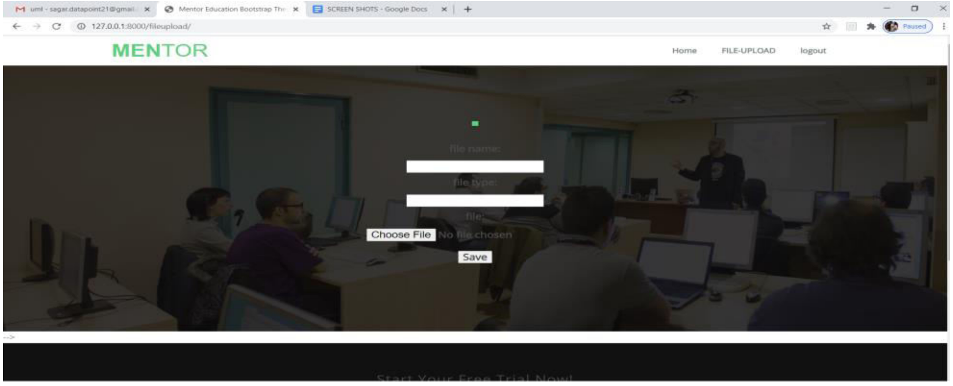
Fig 5 : User Home


uml - sagar.datapoint21@gmail
Mentor Education Bootstrap Th
SCREEN SHOTS - Google Docs
MENTOR
Home
FILE-UPLOAD
logout
Paused
Start Your Free Trial Now!

127.0.0.1:8000/managerloginchecK/

welcome manager-page.

Fig 6 : Manager Home


uml - sagar.datapoint21@gmail
Mentor Education Bootstrap Th
SCREEN SHOTS - Google Docs
MENTOR
Home
FILE-UPLOAD
logout
Paused
Choose File
Save
Start Your Free Trial Now!

127.0.0.1:8000/fileupload/

file name:

file type:

file:

No file chosen

Fig 7: File Upload

7.CONCLUSION AND FUTURE SCOPE

Search engines are very useful for finding out more relevant URLs for given keywords. Due to this, user time is reduced for searching the relevant web page. For this, Accuracy is a very important factor. From the above observation, it can be concluded that XG Boost is better in terms of accuracy than SVM and ANN. Thus, Search engines built using XG Boost and PageRank algorithms will give better accuracy.

FUTURE SCOPE

Machine learning enhance search relevance, personalized results, it can also improve voice based searches, complex query understanding ,real time updates. Ethical AI, and machine learning can aid in understanding user internet. however challenges like data privacy etc.

REFERENCES

[1] Manika Dutta, K. L. Bansal, "A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)", International Journal on Recent and Innovation Trends in Computing and Communication, 2016.

[2] Gunjan H. Agre, Nikita V.Mahajan, "Keyword Focused Web Crawler", International Conference on Electronic and Communication Systems, IEEE, 2015.

[3] Tuhena Sen, Dev Kumar Chaudhary, "Contrastive Study of Simple PageRank, HITS and Weighted PageRank Algorithms: Review", International Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017.

[4] Michael Chau, Hsinchun Chen, "A machine learning approach to web page filtering using content and structure analysis", Decision Support Systems 44 (2008) 482–494,scienceDirect,2008.

[5] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, "Comparative Study of Page Rank and Weighted Page Rank Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, February 2014.

[6] K. R. Srinath, "Page Ranking Algorithms – A Comparison", International Research Journal of Engineering and Technology (IRJET), Dec2017.

[7] S. Prabha, K. Duraiswamy, J. Indhumathi, "Comparative Analysis of Different Page Ranking Algorithms", International Journal of Computer and Information Engineering, 2014.

[8] Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering, 2010.